December 2007

Identity Disclosure Protection: A Data Reconstruction Approach for Preserving Privacy in Data Mining

Dan Zhu
Iowa State University

Shuning Wu
Iowa State University

Xiao-Bai Li
University of Massachusetts Lowell

IDENTITY DISCLOSURE PROTECTION: A DATA RECONSTRUCTION APPROACH FOR PRESERVING PRIVACY IN DATA MINING

Dan Zhu
College of Business
Iowa State University
Ames, IA 50011, U.S.A.
dzhu@iastate.edu

Xiao-Bai Li
College of Management
University of Massachusetts Lowell
Lowell, MA 01854, U.S.A.
xiaobai_li@uml.edu

Shuning Wu
College of Engineering
Iowa State University
Ames, IA 50011, U.S.A.
wushuning@gmail.com

Abstract

Identity disclosure is one of the most serious privacy concerns in today's information age. A well-know method for protecting identity disclosure is k-anonymity. A dataset provides k-anonymity protection if the information for each individual in the dataset cannot be distinguished from at least k – 1 individuals whose information also appears in the dataset. There is a flaw in k-anonymity that would still allow an intruder to discern the confidential information of individuals in the anonymized data. To overcome this problem, we propose a data reconstruction approach to achieve k-anonymity protection in predictive data mining. In this approach, the potentially identifying attributes are first masked using aggregation (for numeric data) and swapping (for nominal data). A genetic algorithm technique is then applied to the masked data to find a good subset of it. This subset is then replicated to form the released dataset that satisfies the k-anonymity constraint.

Keywords: Privacy, identity disclosure, *k*-anonymity, data mining, genetic algorithm

Introduction

Data-mining technologies have enabled organizations to extract useful knowledge from data in order to better understand and serve their customers, and to gain competitive advantages. While successful business applications of data mining are encouraging, there are increasing concerns about compromising the privacy of personal information. A survey by Time/CNN (Greengard 1996) revealed that 93% of respondents believed companies selling personal data should be required to gain permission from the individuals whose information are being shared. In another study (Culnan 1993), more than 70% of participants responded negatively to questions related to the secondary use of private information. Concern about privacy threats has caused data quality and integrity to deteriorate. According to a study by Teltzrow and Kobsa (2004), 82% of online users have refused to give personal information and 34% have lied when asked about their personal habits and preferences.

This study deals with the conflict between privacy and data mining. Organizations that use their customers' records in data-mining activities are obligated to take actions to protect the identities of the individuals involved. It has been demonstrated that personal identities cannot be adequately protected by simply removing identity attributes from released data. There has been extensive research in the area of statistical databases (SDBs) on how to protect individuals' sensitive data when providing summary statistical information. The privacy issue arises in SDBs when summary statistics are derived on very few individuals' data. In this case, releasing the summary statistics may result in disclosing confidential data. The methods for preventing such disclosure can be broadly classified into two categories: (i) query restriction, which prohibits queries that would reveal confidential data, and (ii) data perturbation, which alters individual data in a way such that the summary statistics remain approximately the same. In general, both methods have been extensively investigated and employed (Adam and Workman 1989; Chowdhury et al. 1999; Garfinkel et al. 2002; Sarathy and Muralidhar 2002).

Problems in data mining are somewhat different from those in SDBs. A data-mining task, such as classification or numeric prediction, requires working on individual records contained in a dataset. As a result, query restriction is no longer applicable and data perturbation becomes the primary approach for privacy protection in data mining. Further, predictive data mining essentially relies on discovering relationships between data attributes. Preserving such relationships may not be consistent with preserving summary statistics. Researchers in the data-mining community have proposed various methods to resolve the conflict between data mining and privacy protection. Agrawal and Srikant (2000) considered building a decision tree classifier from data where the confidential values have been perturbed. Evfimievski et al. (2002) presented a framework for mining association rules from transaction data that have been randomized due to privacy concerns. Verykios et al. (2004) investigated the risk of disclosing sensitive rules in data and proposed a set of perturbation algorithms for hiding the sensitive rules. Clifton et al. (2002) discussed techniques for preserving privacy in distributed data mining.

A well-known method for privacy protection, called *k*-anonymity, was recently proposed by Samarati (2001) and Sweeney (2002). The basic idea is to anonymize the data such that each individual cannot be distinguished from a group of other individuals in the data. The method has gained increasing popularity in privacy-preserving data mining. However, the *k*-anonymity approach would, in some circumstances, still allow a data intruder to disclose the individual confidential information in the *k*-anonymized data. To overcome this problem, we propose a data reconstruction approach to achieve *k*-anonymity protection in predictive data mining. In this approach, the potentially identifying attributes are first masked using aggregation (for numeric data) and swapping (for nominal data), without considering the *k*-anonymity constraint. A genetic algorithm technique is then applied to the masked data to find a good subset of it. This subset is then replicated to form the released dataset that satisfies the *k*-anonymity constraint. An experimental study is conducted to show the effectiveness of the proposed method. The proposed approach has broad implications in providing design guidelines for business to implement effective strategies to protect the security and privacy of their customers.

Identity Disclosure Problem

A common practice for protecting identity disclosure is to remove identity related attributes from released data. Sweeney (2002) demonstrated that this is not adequate in protecting personal identities. In fact, the author showed that 87% of the population in the United States can be uniquely identified using three demographic attributes: gender, date of birth, and 5-digit zip code. These attributes are normally not considered identity attributes. However, since they can potentially be used to uniquely identify a record, they are called **quasi-identifiers** (QIs). The *k*-

anonymity technique was proposed to address related identity disclosure problems (Samarati 2001; Sweeney 2002). A dataset provides ***k*-anonymity** protection if the values of the QI attributes for any individual matches those of at least $k - 1$ other individuals in the same dataset. The anonymity is achieved by generalization and suppression of the QI values. With *k*-anonymity, individual identities are better protected. However, as indicated by Machanavajjhala et al. (2006), it is still likely for an intruder to disclose the confidential information of individuals in the *k*-anonymized data. The following hypothetical example demonstrates the problem.

Table 1(a) shows a complete list of 12 patients administered at a hospital in a year for a sensitive disease. The test result is confidential. To protect privacy, the identity related attributes, such as name and address, were removed from the dataset. Knowing they were protected in this way, the patients authorized the hospital to share the data with related professionals and organizations for medical research purposes. However, it would not be difficult for an intruder to find the test results of a patient if he had some knowledge about the patient's age and martial status (quasi-identifiers). Assume, for example, a medical school student, Allen, acquired this dataset from the hospital. If he knew that a 35-year-old, married classmate took this test at the hospital during the year, he can effectively identify his classmate as patient #7, who had a positive test result. Suppose Allen knew that one of his friends, aged 45 and divorced, was also in the list. Then he could also easily find that his friend was patient #12, who also had a positive test result.

Table 1. An Illustrative Example

(a) Original Patient Data

No.	Age	Marital Status	Blood Pressure	Blood Type	Test Result
1	26	Never married	75/120	O	Negative
2	27	Never married	86/133	A	Positive
3	27	Never married	70/115	O	Negative
4	28	Never married	90/140	B	Negative
5	32	Married	80/135	AB	Negative
6	34	Married	83/147	O	Positive
7	35	Married	95/144	A	Positive
8	35	Divorced	65/112	O	Negative
9	40	Widow	78/136	A	Positive
10	43	Married	110/155	AB	Positive
11	45	Married	100/150	O	Positive
12	45	Divorced	84/135	A	Positive

(b) *k*-Anonymized Patient Data ($k = 4$)

No.	Age	Marital Status	Blood Pressure	Blood Type	Test Result
1	20-29	Never married	75/120	O	Negative
2	20-29	Never married	86/133	A	Positive
3	20-29	Never married	70/115	O	Negative
4	20-29	Never married	90/140	B	Negative
5	30-39	Married	80/135	AB	Negative
6	30-39	Married	83/147	O	Positive
7	30-39	Married	95/144	A	Positive
8	30-39	Married	65/112	O	Negative
9	40-49	Married	78/136	A	Positive
10	40-49	Married	110/155	AB	Positive
11	40-49	Married	100/150	O	Positive
12	40-49	Married	84/135	A	Positive

The *k*-anonymity technique can help protect against such identity disclosure problems. Table 1(b) shows the anonymized dataset released by the hospital. The generalization method was applied to the original data where the age values were grouped into three intervals and marital status values were combined into two groups (with Married representing three original categories: Married, Divorced and Widow). From this dataset, Allen can no longer identify his classmate (#7) or the classmate's test result. As far as his other friend (#12) is concerned, however, Allen is still able to access confidential information. Although he cannot identify which record is his friend's, he still knows that his friend has a positive test result, since all of the four records in the group containing his friend's record have the same test result.

The example above demonstrates that it is still quite possible for a data intruder to disclose the confidential information of an individual in the *k*-anonymized data. *K*-anonymity protects identity disclosure by generalizing different but similar QI attribute values into the same value. The new values produced by the generalization operation are still correct with respect to the generalized categories. Since confidential values (e.g., test result) remain unchanged in *k*-anonymity, individuals in a group are subject to high disclosure risk if their confidential values in the group are the same. To overcome this problem, Machanavajjhala et al. (2006) proposed a new privacy principal, called *l*-diversity, which requires, in addition to *k*-anonymity, that the confidential attribute should include at least *l* "well-represented" values in the anonymized data. This additional constraint can sometimes be hard to satisfy and usually causes much larger group sizes.

Another drawback with the *k*-anonymity approach is that it destroys the univariate statistical properties of the QI attributes, which are very important in statistical and data warehousing applications. This problem is due to the use of generalization and suppression methods: generalization creates new categorical values instead of keeping the original categorical values in the data, while suppression results in skewed distributions (partial suppression) or no univariate information at all (full suppression) for the QI attributes. This loss of univariate information also exists in other *k*-anonymity based techniques such as *l*-diversity. The problem becomes worse when the technique is geared towards specific data-mining algorithms, such as that proposed in Friedman et al. (2007).

The data reconstruction approach we propose addresses both the privacy protection and information loss problems mentioned above. Our proposed method masks the QI attributes by aggregating numeric values and swapping nominal values. Aggregation and swapping operations differ from generalization in that the aggregated and swapped values are "faked" values (as opposed to "correct" values produced by generalization). As a result, the proposed approach provides a better protection against identity disclosure. In addition, aggregation preserves approximately some important numeric univariate statistics (e.g., mean), while swapping completely preserves the univariate (frequency) distribution of a nominal attribute.

The Data Reconstruction Approach

This study deals with privacy protection problem in the context of predictive data mining. We focus our approach on classification analysis, which is a common data-mining task. The basic idea of our approach also applies to the other predictive data-mining tasks such as numerical prediction (regression). We do not, however, target unsupervised learning problems such as clustering and association rules mining [see Aggarwal et al. (2006) and Friedman et al. (2007) for example studies in these areas]. The objective of our approach is to preserve classification accuracy while achieving *k*-anonymity. We are interested in situations where the class attribute is confidential (if the confidential attribute is a non-class attribute, it should be somewhat easier to accomplish the above objective). In this setting, there are three types of attributes:

- *Confidential* attribute, which contain private information that an individual typically does not want revealed, such as test result in the illustrative example. In *k*-anonymity, a confidential attribute will not be masked.
- *Quasi-identifiers*, which can be obtained from other sources and then used to identify an individual, such as age and marital status in the example. Quasi-identifiers will be masked in *k*-anonymity.
- *Non-QI* attributes, which is unlikely to be known by an intruder, such as blood pressure and blood type in the example. These attributes will not be changed in *k*-anonymity.

We first apply numeric value aggregation to numeric QI attributes and then nominal value swapping to nominal QI attributes.

Numeric Value Aggregation

For each numeric QI attribute, a supervised discretization method (Fayyad and Irani 1992) is used to divide the numeric values into groups. The goal of this method is to preserve the relationships between the class attribute and the numeric attributes after discretization. The method recursively splits an attribute to minimize the class entropy and uses a minimum description length criterion to determine when to stop. The algorithm evaluates the information gain on each of the potential cut points, and chooses the one with the maximum value to split. This process is repeated recursively until a stopping criterion is reached (e.g., the subset contains a single class only). Groups are formed based on the cut points, and the value of a numeric attribute is subsequently set for each instance as the median value in corresponding group. For example, in Table 1(a), when Age is considered as a QI attribute, its values will be divided into two groups based on this algorithm:

Group 1: age ≤ 34, with median = 28;
Group 2: age > 34, with median = 40.

Then the age values are set to 28 for the first six instances, and 40 for the last six instances.

Nominal Value Swapping

For each nominal QI attribute, we use a data swapping method, based on Reiss (1984), to mask the value of the attribute. With this method, a part of the current values of the QI attribute and the class attribute are replaced with new values such that the lower-order statistical distributions of the masked data will be close to those of the original data. We apply a second-order feedback algorithm as described in Reiss (1984), which computes the second-order frequency distributions of the original data, and swaps the data such that the new data have approximately the same distributions. In Reiss (1984), the data are assumed to be 0-1 valued; i.e., the number of different values for each nominal attribute $r = 2$. We extend the algorithm to arbitrary r to handle general multi-category data.

For classification problems, it is import to maintain the 2nd-order statistics between the class attribute and each of the QI attributes, because in many classification techniques, such as decision trees and naïve Bayes method, classification models are built based on such 2nd-order statistics. Let Y be the class attributes, which has M categories. Let $X_1,...,X_Q$ be the Q nominal QI attributes, including the discretized numerical QI attributes. Let N be the total number of instances of the dataset. The algorithm is described below:

1. Compute the 1st-order frequency tables $F_1(X_q)$ $(q = 1,...,Q)$ and the 2nd-order frequency tables $F_2(X_q, Y)$ $(q = 1,...,Q)$ using the original data.

2. For $q = 1, \ldots, Q$, perform the following swapping procedure:

Find a masked $N \times 2$ dataset, $D = \{Z, Y\}$, where Z corresponds to the masked values of X_q. We want the 1st- and 2nd-order frequency distributions of D to be as close to $F_1(X_q)$ and $F_2(X_q, Y)$ as possible. This is implemented as below:

FOR $i = 1$ TO N DO

Choose $(i,\ f_1(j),\ x_{qj})$, $\forall\ j = 1,...,J_q$,

Choose $(i,\ f_2(j,m),\ x_{qj})$, $\forall\ j = 1,...,J_q$ and $m = 1,...,M$.

END

where x_{qj} is the jth category of attribute X_q, J_q is the number of categories in X_q, and M is the number of categories in Y. *Choose* (i, f, x) sets the ith instance of Z as $z_i = x$ with probability f. And $f_1(j)$ and $f_2(j,m)$ are defined respectively as:

$$f_1(j) = \frac{F_1(X_q = x_{qj})}{N}, \tag{1}$$

where $F_1(X_q = x_{qj})$ is the count for attribute value x_{qj}; and

$$f_2(j,m) = \begin{cases} \dfrac{F_2(X_q = z_i, Y = y_m)}{F_1(X_q = z_i)} & \text{if } F_1(X_q = z_i) \neq 0 \\ 1/J_q & otherwise \end{cases} \tag{2}$$

where y_m is the mth category of the class attribute Y, and $F_2(X_q = z_i, Y = y_m)$ is the count when $X_q = z_i$ and $Y = y_m$. At the end of each *Choose*() iteration, the counts F_1 and F_2 are updated and the algorithm will incorporate such feedback.

We should point out that this swapping algorithm aims at preserving the 2nd-order statistics between the class attribute and each QI attribute, but it makes no effort to preserve such statistics between a QI attribute and a non-QI (or another QI) attribute. If the released data will also be used to study such relationships, the above algorithm can be modified by adding additional iterations where the class attribute Y is replaced with a concerned QI or non-QI attribute. Modifications can also be made to preserve higher-order statistics. For instance, if we want to study the joint impact of Age and Blood Type on Test Result in the above example, we can create a compound attribute (called, say, "Age × Blood Type") that includes all possible combinations of aggregated Age and Blood Type values as its values. The above algorithm can then be applied in terms of Test Result and the compound attribute. The computational cost will increase when attempting to preserve more and higher-order relationships, however.

Genetic Algorithm Based Instance Selection

Next, we consider masking the data to achieve k-anonymity. The basic idea is to first apply an instance selection technique to find a good subset, S, of n records, where $n = N / k$, and then replicate the QI attribute values of each record in S for $k - 1$ times to get a full dataset of size N, which satisfies k-anonymity.

The instance selection technique we use is based on genetic algorithms (GAs) (Goldberg 1989). The process of natural evolution and genetics has been studied by computing and biology scientists and enormous progress has been made over the past two decades. Genetic algorithms have been applied to a variety of applications, including design, control, scheduling and other optimization problems. Prior research in instance selection has shown that evolutionary-based techniques like GAs outperform traditional methods such as random sampling and nearest neighbour search. GAs typically result in higher classification accuracy and smaller subset size (Reeves and Bush 2001; Ishibuchi et al. 2001; Cano et al. 2003).

A genetic algorithm (GA) is a search technique to find approximate solutions to optimization and search problems. GAs are a particular class of evolutionary algorithms that use techniques inspired by evolutionary biology such as inheritance, mutation, natural selection, and recombination (or crossover). They are typically implemented as a computer simulation in which a population of abstract representations of candidate solutions (called *chromosomes*) to an optimization problem evolves toward better solutions. The evolution starts from a population of completely random chromosomes and takes place in generations. In each generation, the fitness of the whole population is evaluated, multiple chromosomes are stochastically selected from the current population (based on their fitness), modified (mutated or recombined) to form a new population, which becomes current in the next iteration of the algorithm.

The whole instance selection process is illustrated in Figure 1. To begin, the original data set is randomly divided into two parts: data set T^* and test set D^*. T^* is then sampled with replacement $|T^*|$ times, each taking one instance, to generate the training set T. The instances that are not selected form an independent validation set $D = T^* \setminus T$. (Note that D is used to validate the classification models in the process of GA algorithm, while D^* is used to evaluate the performance of the models built on the final GA outputs.)

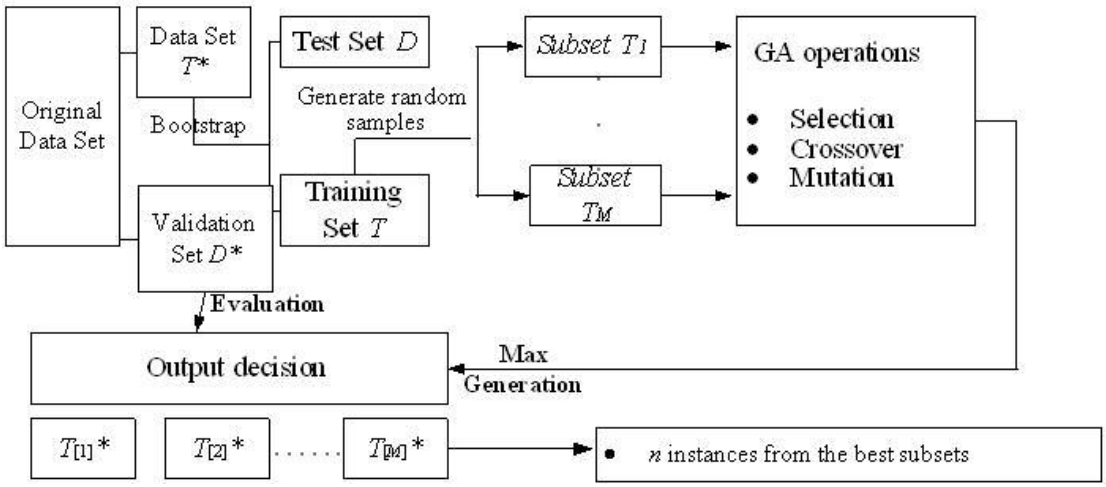


Figure 1. GA-based instance selection

Fitness function and solution representation

Let $T = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ be the training data set. Our objective is to select a subset $S \subset T$ such that the classification model $\psi(S)$ induced on this subset is able to maintain adequate prediction accuracy compared to the model $\psi(T)$ induced on the entire training set. We hence define the fitness function as follows:

$$f(S) = -\log(1 - \hat{e}(\psi(S))) \tag{3}$$

where $\hat{e}(\psi)$ is the estimate of the error rate of the classification model ψ. The estimation is done using a bootstrapping approach. We know that the chance that a particular instance is not selected for the training set T is $(1-\frac{1}{n})^n \approx e^{-1} = 0.368$, where e is the base of natural logarithms, 2.7183. Thus for a reasonably large data set, the validation set D will contain about 36.8% of the instances, and the training set T will contain about 63.2% of them. A classification model $\psi(S)$ is built based on a subsets and its estimated error rate is evaluated using sets T and D, as below (Witten and Frank 2005):

$$\hat{e}(\psi(S)) = 0.632 \cdot e_D(\psi(S)) + 0.368 \cdot e_T(\psi(S)) \tag{4}$$

where $e_D(\psi(S))$ is the error when the model $\psi(S)$ is applied to set D, and $e_T(\psi(S))$ is the error when it is applied to set T.

The optimization problem is thus to find the subset that minimizes (4). We use a GA implementation to find a heuristic solution to this problem. The solution space is defined in terms of chromosomes, each of which represents a subset of instances in T. Let g_i denote the position of the ith instance in T. The chromosomal unit of each subset is defined as a vector $C = [g_1, g_2, \cdots\cdots, g_V]$ of integers, where V is the number of instances in a subset and represents the length of the chromosomal unit.

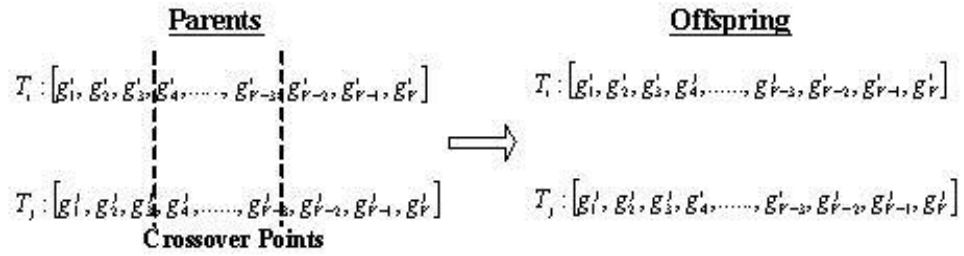
GA operations and heuristic solution

The GA search starts with an initial population $P_0 = \{C_1^{(0)}, C_2^{(0)}, ..., C_k^{(0)}\}$ of chromosomes that is selected by dividing T into k subsets by sampling with replacement from T. Each chromosome has a size of $V = \lfloor |T|/k \rfloor$. Starting with this initial population, the usual GA operations of selection, crossover, and mutation are applied to improve the population. These operations are described as follows.

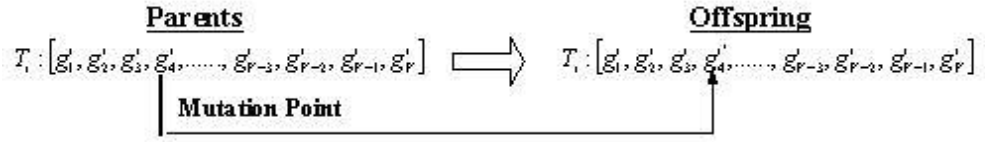
For the selection, in the *h*th step of the GA search, the current population P_h is ranked according to the fitness,

$$f\left(S\left(C_{[1]}^{(h)}\right)\right) \geq f\left(S\left(C_{[2]}^{(h)}\right)\right) \geq \ldots \geq f\left(S\left(C_{[k]}^{(h)}\right)\right) \tag{5}$$

where $S\left(C_{[j]}^{(h)}\right) = \left\{\mathbf{x}_i : i \in C_{[j]}^{(h)}\right\}$ is the subset corresponding to chromosome $C_{[j]}^{(h)}$, j = 1, 2,…, k. The $(1-c)k$ fittest ones are selected into the next generation, where c is the crossover rate. The crossover operator probabilistically selects $ck/2$ pairs from P_h, randomly chooses two crossover sections (of the same size) from two chromosomes, and then swaps the crossover sections between the two chromosomes (see Figure 2(a)). In mutation, an element g_i in the chromosomal unit is chosen with some probability, and a new random number g_i' is generated uniformly from {1, 2, …, $|T|$} to replace g_i (see Figure 2(b)).



(a)



(b)

Figure 2: Operations of the Genetic Algorithm: (a) Crossover (b) Mutation

The GA operations are repeated for a specified number of G generations, resulting in a final population $P_G = \left\{C_{[1]}^{(G)}, C_{[2]}^{(G)}, \ldots, C_{[k]}^{(G)}\right\}$, which is ranked as in (5) above. The heuristic solution to the best subset is then obtained by selecting n instances that are contained in the top j chromosomes in P_G; that is,

$$S^* = \left\{\mathbf{x}_i : i \in C_{[1]}^{(G)} \bigcup C_{[2]}^{(G)} \ldots \bigcup C_{[j]}^{(G)} \text{ and } |S^*| = n\right\}. \tag{6}$$

K-Anonymity Procedure

Once we have obtained S^*, the final step is to create k-anonymity for the original dataset based on S^*. For each instance i in S^*, this is done by finding the $k-1$ nearest instances to i from the original set, and then changing the QI attribute values of these instances to those of instance i. The pseudo code for this approach is as following:

FOR instance i = 1 TO n in S^* DO

- Compute the Euclidean distance between instance i and each instance in the original set, where numeric data are normalized to [0, 1], and distance between two nominal values is defined as zero if they are the same, and one otherwise.
- Find $(k-1)$ instances having the smallest distance to instance i.
- Replace the values of the QI attributes for these $(k-1)$ instances with the values of the corresponding QI attributes of instance i.

END

Computational Complexity

The time complexity for the numeric value discretization procedure is of $O(N \log N)$ (Fayyad and Irani 1992). For nominal value swapping, the computational cost is also low, since only the 1st-order statistics and a part of the 2nd-order statistics are involved. It is clear from the swapping algorithm that the time complexity is of $O(M \sum_{q=1}^{Q} J_q)$, where M is the number of categories in the confidential attribute and J_q is the number of categories in the qth QI attribute. So, $M \sum_{q=1}^{Q} J_q$ is the total number of nominal value combinations between the confidential attribute and QI attributes (after aggregation). This quantity is typically smaller than N in a large classification problem.

A GA generation essentially involves building k decision trees, each based on a subset of size $V = \lfloor |T|/k \rfloor$. So, the time complexity of the whole GA procedure is of $O(GkV \log V)$, which is smaller than $O(GN \log N)$, where the number of generations G can be set by the user to a reasonable level. The convergence of GAs has been analyzed by numerous researchers (Ding and Yu 2005; Rudolph 1994; Suzuki 1995). Empirical studies by Ishibuchi et al. (2001), Reeves and Bush (2001), and Li and Varghese (2007) have also shown the GA's convergence behaviour in instance selection problems.

Finally, it is clear that the procedure for creating k-anonymized dataset from S^* is of order $O(nN)$, where n and N are the size of S^* and the original dataset, respectively. To sum up, the worst-case complexity for the entire proposed algorithm is of order $O(N \log N) + O(M \sum_{q=1}^{Q} J_q) + O(GN \log N) + O(nN)$, which should be somewhere between $O(N \log N)$ and $O(N^2)$. Clearly, this time complexity is comparable to those of major data mining algorithms, and the proposed algorithm is scalable to large data size.

An Illustrative Example

In this section, we demonstrate our approach using the example data in Table 1(a). As mentioned earlier, Test Result is the confidential attribute in this dataset. Age and Marital Status are the QI attributes, and Blood Pressure and Blood Type are non-QI attributes. In k-anonymity, the QI attributes are masked while the other attributes are unchanged. To illustrate, let $k = 2$.

Step 1. Apply discretization procedure to convert numerical Age values into discretized values (labeled as Age2). The attribute is discretized into two values: 28 and 40.

Table 2. Discretized and Swapped Data

No.	Age2	Marital Status	Blood Pressure	Blood Type	Test Result
1	28	Never married	75/120	O	Negative
2	40	Married	86/133	A	Positive
3	28	Never married	70/115	O	Negative
4	28	Never married	90/140	B	Negative
5	28	Married	80/135	AB	Negative
6	40	Married	83/147	O	Positive
7	40	Divorced	95/144	A	Positive
8	40	Widow	65/112	O	Negative
9	40	Divorced	78/136	A	Positive
10	40	Married	110/155	AB	Positive
11	28	Never Married	100/150	O	Positive
12	28	Married	84/135	A	Positive

Table 3. The 2nd-Order Statistics before and after Swapping

Age2	Test Result	Original Counts	Counts after Swapping	Marital Status	Test Result	Original Counts	Counts after Swapping
28	Negative	4	4	Never Married	Negative	3	3
28	Positive	2	2	Never Married	Positive	1	1
40	Negative	1	1	Married	Negative	1	1
40	Positive	5	5	Married	Positive	4	4
				Divorced	Negative	1	0
				Divorced	Positive	1	2
				Widow	Negative	0	1
				Widow	Positive	1	0

Step 2. Apply the Reiss 2nd-order swapping procedure to swap the Marital Status and Age2 attribute values. The 2nd-order statistics that we want to preserve as much as possible is the count for the categorical combinations between the class attribute Test Result and Marital Status, and that between Test Result and Age2. The results after discretization and swapping are shown in Table 2. The 2nd-order statistics before and after swapping are given in Table 3.

Step 3. GA-based procedure is applied and the resulting subset of six instances is shown in Table 4.

Table 4. Results from GA-based Selection

No.	Age2	Marital Status	Blood Pressure	Blood Type	Test Result
1	28	Never married	75/120	O	Negative
3	28	Never married	70/115	O	Negative
5	28	Married	80/135	AB	Negative
6	40	Married	83/147	O	Positive
7	40	Divorced	95/144	A	Positive
10	40	Married	110/155	AB	Positive

Table 5. *k*-Anonymized Results

No.	Closest Instance	Age	Marital Status	Blood Pressure	Blood Type	Test Result
1		28	Never married	75/120	O	Negative
2	7	40	Divorced	86/133	A	Positive
3		28	Never married	70/115	O	Negative
4	1	28	Never married	90/140	B	Negative
5		28	Married	80/135	AB	Negative
6		40	Married	83/147	O	Positive
7		40	Divorced	95/144	A	Positive
8	3	28	Never married	65/112	O	Negative
9	10	40	Married	78/136	A	Positive
10		40	Married	110/155	AB	Positive
11	6	40	Married	100/150	O	Positive
12	5	28	Married	84/135	A	Positive

Step 4. Perform *k*-anonymity using partial duplication. Here all of the non-confidential attributes are used in calculating the Euclidean distance to find the $k - 1$ closest instances, while only the QI attribute values (i.e., Age2 and Marital Status) are used for duplication. The k-anonymized data are shown in Table 5.

Computational Experiments and Results

A set of numerical experiments were conducted using two real-world datasets. Both datasets were taken from the Machine Learning Repository of the University of California at Irvine (Hettich and Bay 1999). The characteristics of these two datasets are described in Table 6.

Table 6. Test Datasets

	Number of Instances	Number of Attributes	Number of Classes
Diabetes	768	9	2
German credit	1000	21	2

The first dataset, Diabetes, contains 768 instances of patient information, with 9 numerical and nominal attributes, including diagnostic result, number of times pregnant, age, and a few lab test measures. Diagnostic result was considered as the confidential (class) attribute. Number of times pregnant and age were considered as the QI attributes subject to masking. Those lab test measures were non-QI attributes. The second dataset, German Credit, consists of 21 numerical and nominal attributes, representing credit rating, age, gender with marital status, years of employment, housing type, job type, phone status, foreign worker indicator, and a number of attributes related to the customer's account. Credit rating (categorical) was considered as confidential (class) attribute. The other seven attributes mentioned above were considered as the QI attributes. Those account related attributes were non-QI attributes.

The effectiveness of the proposed method is evaluated using two popular classification techniques, the C4.5 decision tree classifier (Quinlan, 1993) and support vector machine (Vapnik 1995). Each dataset was randomly divided into two parts: approximately 75% for data set T^*, and 25% for test set D^*. The data sets serve as the original set for masking, while the test sets are not masked. Furthermore, to show the effectiveness of our method for privacy protection, a disclosure risk measure called record linkage (Pagliuca and Seri 1999) is used. Record linkage measures disclosure risk using the Euclidean distances between records in the masked dataset and those in the original set. It is primarily used for numeric data, which are normalized to [0, 1]. For nominal data, a common practice is to assign a distance of zero if the corresponding attribute values of the two records are the same; otherwise, the distance is one. A record in the masked set is said to be "linked" if the closest record in the original set is indeed the corresponding unmasked record. A record in the masked set is "second closely linked" if the second closest record in the original set is the corresponding one. Then the record linkage measure is defined as the percentage of records that are either "linked" or "second closely linked".

The anonymity degree k is varied with three values: $k = 2, 4, 6$. For each dataset, the proposed method was run ten times. The average results of the classification accuracies were then reported. The results on the record linkage ratio (RL ratio) were also reported for comparison. In order to examine the effect of GA-based *k*-anonymity procedure, we also performed the experiments based on the data after aggregation and swapping, but before instance selection procedure. Tables 7 and 8 show the results of these experiments.

The original accuracies, the accuracies and RL ratios before the GA instance selection (but after the aggregation and swapping) are presented in the left-most columns for each data set. It can be seen from Tables 7 and 8 that aggregation and swapping preserve the classification accuracies very well. However, the RL ratios after aggregation and swapping are fairly large, particularly in the second (Diabetes) data set, which indicates that the disclosure risks are still high.

Table 7. Results of Experiments for German Credit Data

Original **C4.5** accuracy: 70.5%	$k = 2$		$k = 4$		$k = 6$	
	Accuracy	RL Ratio	Accuracy	RL Ratio	Accuracy	RL Ratio
Accuracy before GA: 70.4% RL ratio before GA: 9.3%	69.8%	0.80%	67.6%	0.27%	63.6%	0.01%
Original **SVM** accuracy: 75.1%	$k = 2$		$k = 4$		$k = 6$	
	Accuracy	RL Ratio	Accuracy	RL Ratio	Accuracy	RL Ratio
Accuracy before GA: 75.0% RL ratio before GA: 9.2%	74.0%	0.93%	68.4%	0.80%	65.6%	0.67%

Table 8. Results of Experiments for Diabetes Data

Original **C4.5** accuracy: 75.1%	$k = 2$		$k = 4$		$k = 6$	
	Accuracy	RL Ratio	Accuracy	RL Ratio	Accuracy	RL Ratio
Accuracy before GA: 73.4% RL ratio before GA: 39.76%	71.9%	0.17%	70.4%	0.09%	68.8%	0.01%
Original **SVM** accuracy: 77.0%	$k = 2$		$k = 4$		$k = 6$	
	Accuracy	RL Ratio	Accuracy	RL Ratio	Accuracy	RL Ratio
Accuracy before GA: 75.0% RL ratio before GA: 39.58%	75.0%	1.04%	73.3%	0.35%	71.3%	0.17%

In order to examine the convergence of our GA-based instance selection approach, we record the classification accuracy for each different number of GA generations. The results with decision trees for $k = 4$ are illustrated in Figure 3. It appears that accuracies increase steadily till about the generation 50 and then level off afterwards. The number of GA generations G is thus set to 50.This convergence rate is very fast and actually faster than that of many other GA applications, although similar situations have also been observed in the other instance selection studies using GAs (Li and Jacob 2007).

It is observed from Tables 7 and 8 that as k increases, the classification accuracies decrease, as well as the record linkage ratios. More importantly, although decision trees based on the data after GA-based instance selection produce somewhat lower classification accuracy, record linkage values based on these data drop much more significantly. This indicates that the GA-based k-anonymity significantly reduce the disclosure risk while still maintaining reasonable data quality. The results on support vector machine (SVM) indicate that the proposed method works not only for decision trees, but also for other classification methods.

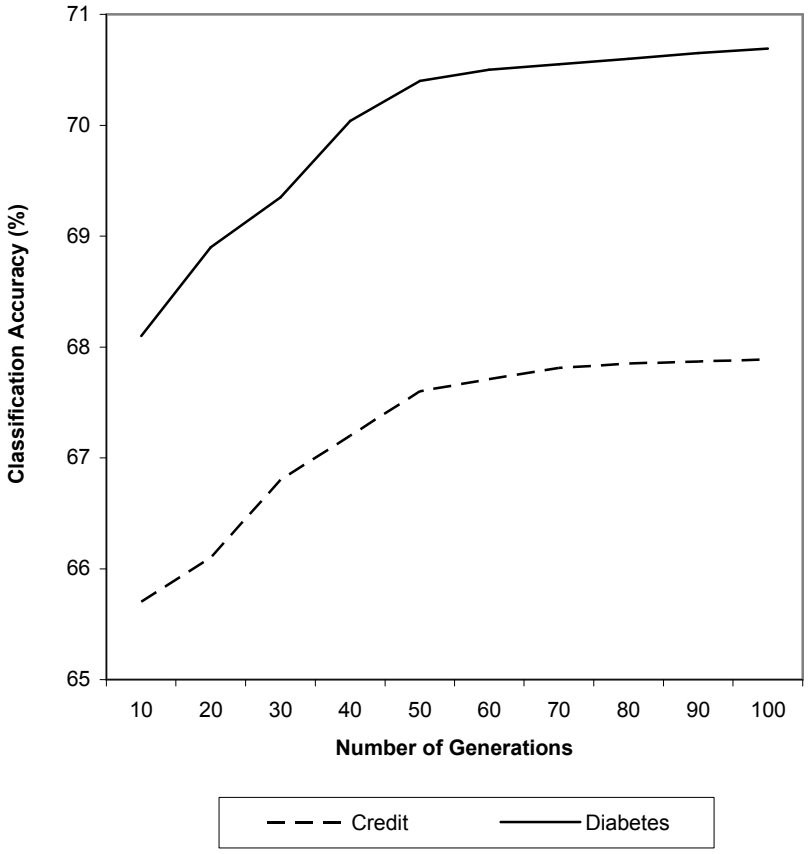


Figure 3. GA Generations vs. Classification Accuracies

Conclusion and Discussion

This paper presents a novel instance selection method based on genetic algorithm for identity disclosure protection. We introduce a data reconstruction approach to achieve *k*-anonymity protection in privacy-preserving data mining. An experimental study is conducted to show the effectiveness of the proposed method. Our empirical evaluation results indicate that our proposed approach can lead to significantly improved performance. The insights gained from this study can help business make effective decisions on privacy protection in data mining.

Our work illustrates the usefulness of using instance selection for privacy protection, and the effectiveness of using genetic algorithm for obtaining heuristic solutions to this problem. Future research will take into account more complicated situations, and in particular characterize dataset where this approach is most likely to work well. In particular, we will consider how such parameters as number of instances, number of class values, and number of attributes influence the performance of the algorithm.

In a classification problem, there is only one class attribute. By designating the class attributes confidential and non-class attribute non-confidential, we have implicitly assumed that there is only one confidential attribute in the data. The proposed method can be extended to handle multiple categorical confidential attributes. In this situation, we can consider all confidential attributes together as one compound attribute. Suppose, for instance, the Marital Status attribute in the earlier example (Table 1) is also confidential. A compound attribute called "Marital Status × Test Result" can be created, which would have eight categories, formed by different combinations of Marital Status and Test Result values. The transformed dataset would have three non-confidential and one (compound) confidential attributes. The proposed method can then be applied to this transformed dataset.

References

Adam, N. R., and Wortmann, J. C. "Security-Control Methods for Statistical Databases: A Comparative Study," *ACM Computing Surveys* (21:4), 1989, pp. 515-556.

Aggarwal, G., Feder, T., Kenthapadi, K., Khuller, S., Panigrahy, R., Thomas, D., and Zhu A. "Achieving Anonymity via Clustering," in *Proceedings of the 25th Symposium on Principles of Database Systems* (*PODS'06*), Chicago, IL, 2006, pp. 153-162.

Agrawal, R., and Srikant, R. "Privacy-Preserving Data Mining," in *Proceedings of 2000 ACM SIGMOD International Conference on Management of Data*, Dallas, TX, 2000, pp. 439-450.

Berndt, E. R. *The Practice of Econometrics*, Addison-Wesley, NY, 1991.

Cano, J. R., Herrera, F., and Lozano, M. "Using Evolutionary Algorithms as Instance Selection for Data Reduction in KDD: An Experimental Study," *IEEE Transactions on Evolutionary Computation* (**7**:6), 2003, pp. 561-575.

Chowdhury, D. S., Duncan, G. T., Krishnan, R., Roehrig, S. F., and Mukherjee, S. "Disclosure Detection in Multivariate Categorical Databases: Auditing Confidentiality Protection through Two New Matrix Operators," *Management Science* (45:12), 1999, pp. 1710-1723.

Clifton, C., Kantarcioglu, M., Vaidya, J., Lin X., and Zhu, M. "Tools for Privacy Preserving Distributed Data Mining," *SIGKDD Explorations* (4:2), 2002, pp. 38-44.

Culnan, M. "'How Did They Get My Name?': An Exploratory Investigation of Consumer Attitudes toward Secondary Information Use," *MIS Quarterly* (17:3), 1993, pp. 341-363.

Ding, L., and Yu. J. "Some Theoretical Results about the Computation Time of Evolutionary Algorithms," in *Proceedings of the 2005 Conference on Genetic and Evolutionary Computation*, Washington, DC, 2005, pp. 1409-1415.

Evfimievski, A., Srikant, R., Agrawal, R., and Gehrke, J. "Privacy Preserving Mining of Association Rules," in *Proceedings of 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Canada, 2002, pp. 217-228.

Fayyad, U. M., and Irani, K. B. "On the Handling of Continuous-Valued Attributes in Decision Tree Generation," *Machine Learning* 8, 1992, pp. 87-102.

Fayyad, U., Piatetsky-Shapiro, G., Smith, P., and Uthurusamy R. *Advances in Knowledge Discovery and Data Mining*. MIT Press, Cambridge, MA, 1996.

Friedman, A., Schuster, A., and Wolff, R. "Providing k-Anonymity in Data Mining," *International Journal on Very Large Data Bases*, Forthcoming 2007.

Galletta, D. MIS Faculty Salary Survey Results. Accessed March 2004, from http://www.pitt.edu/~galletta/salsurv.html.

Garfinkel, R., Gopal, R., and Goes, P. "Privacy Protection of Binary Confidential Data against Deterministic, Stochastic, and Insider Threat," *Management Science* (48:6), 2002, pp. 749-764.

Goldberg, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning,* Kluwer Academic Publishers, Boston, MA, 1989.

Greengard, S. "Privacy: Entitlement or Illusion?" *Personnel Journal* (75:5), 1996, pp. 74-88.

Greenhalgh, D., and Marshall, S. "Convergence Criteria for Genetic Algorithm," *SIAM Journal on Computing* (30:1), 2000, pp. 269 – 282.

Hettich, S. and Bay S. D. The UCI KDD Archive [http://kdd.ics.uci.edu]. University of California, Department of Information and Computer Science, Irvine, CA, 1999.

Ishibuchi, H., Nakashima, T., and Nii, M. "Genetic-Algorithm-Based Instance Selection and Feature Selection," in *Instance Selection and Construction for Data Mining*, Liu, H. and Motoda, H. (Eds.), Kluwer Academic, Norwell, MA, 2001, pp. 96-112.

Li, X.-B., and Jacob, V. S. "Adaptive Data Reduction for Large-Scale Transaction Data," *European Journal of Operational Research*, Forthcoming 2007.

Machanavajjhala, A., Gehrke, J., Kifer, D., and Venkitasubramaniam, M. "*l*-Diversity: Privacy beyond *k*-Anonymity," in *Proceedings of the 22nd IEEE International Conference on Data Engineering (ICDE 2006)*, Atlanta, Georgia, 2006.

Pagliuca, D., and Seri. G. "Some Results of Individual Ranking Method on the System of Enterprise Accounts Annual Survey," Esprit SDC Project, Deliverable MI-3/D2, 1999.

Quinlan, J. R. *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA, 1993.

Reeves, C. R., and Bush, D. R. "Using Genetic Algorithms for Training Data Selection in RBF Networks," in *Instance Selection and Construction for Data Mining*, Liu H. and Motoda, H. (Eds.), Kluwer Academic, Norwell, MA, 2001, pp. 339-356.

Reiss, S. P. Practical data-swapping: the first steps, *ACM Transactions on Database Systems (TODS)*, v.9 n.1, p.20-37, March 1984.

Rudolph, G. "Convergence Analysis of Canonical Genetic Algorithms," *IEEE Transactions on Neural Networks* (5:1), 1994, pp. 96-101.

Samarati, P. "Protecting Respondents' Identities in Microdata Release," *IEEE Transactions on Knowledge and Data Engineering* (13:6), 2001, pp. 1010-1027.

Sarathy, R. and Muralidhar, K. "The Security of Confidential Numerical Data in Databases," *Information Systems Research* (13:4), 2002, pp. 389-403.

Suzuki, J. "A Markov Chain Analysis on Simple Genetic Algorithms," *IEEE Transactions on Systems, Man and Cybernetics* (25:4), 1995, pp. 655-659.

Sweeney, L. "k-Anonymity: A Model for Protecting Privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* (10:5), 2002, pp. 557-570.

Teltzrow, M., and Kobsa, A. "Impacts of User Privacy Preferences on Personalized Systems: A Comparative Study," in *Designing Personalized User Experiences in eCommerce*, Karat, C. M., Blom, J., Karat, J. (Eds.), Kluwer Academic Publishers, Dordrecht, Netherlands, 2004, pp. 315–332.

Vapnik, V.N. *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.

Verykios, V. S., Elmagarmid, A. K., Bertino, E., Saygin, Y., and Dasseni, E. "Association Rule Hiding," *IEEE Transactions on Knowledge and Data Engineering* (16:4), 2004, pp. 434-447.

Witten, I. H., and Frank, E. *Data mining: Practical machine learning tools and techniques*, second edition, Morgan Kaufmann, San Francisco, 2005.